Utilice estos diez principios para guiar la evaluación de la investigación, instan Diana Hicks, Paul Wouters y sus colegas.

Crédito: Ilustración de David Parkins

Los datos se utilizan cada vez más para gobernar la ciencia. Las evaluaciones de la investigación, que antes se hacían a medida y las realizaban colegas, ahora son rutinarias y dependen de las métricas1. El problema es que la evaluación se rige ahora por los datos y no por el juicio. Las métricas han proliferado: normalmente bien intencionadas, no siempre bien informadas, a menudo mal aplicadas. Corremos el riesgo de dañar el sistema con las mismas herramientas diseñadas para mejorarlo, ya que la evaluación es implementada cada vez más por organizaciones sin conocimiento de las buenas prácticas y la interpretación, ni asesoramiento al respecto.

Antes de 2000, existía el Science Citation Index en CD-ROM del Institute for Scientific Information (ISI), utilizado por los expertos para los análisis especializados. En 2002, Thomson Reuters puso en marcha una plataforma web integrada, haciendo que la base de datos Web of Science fuera ampliamente accesible. Se crearon índices de citas competitivos: Scopus de Elsevier (publicado en 2004) y Google Scholar (versión beta publicada en 2004). Se introdujeron herramientas basadas en la web para comparar fácilmente la productividad y el impacto de la investigación institucional, como InCites (que utiliza la Web of Science) y SciVal (que utiliza Scopus), así como un software para analizar los perfiles de citación individuales utilizando Google Scholar (Publish or Perish, publicado en 2007).

En 2005, Jorge Hirsch, físico de la Universidad de California en San Diego, propuso el índice h, popularizando el recuento de citas para investigadores individuales. El interés por el factor de impacto de las revistas no dejó de crecer a partir de 1995 (véase "Obsesión por el factor de impacto").

Últimamente, las métricas relacionadas con el uso social y los comentarios en línea han cobrado impulso: en 2002 se creó F1000Prime, en 2008 Mendeley y en 2011 Altmetric.com (apoyado por Macmillan Science and Education, propietaria de Nature Publishing Group).

Como cienciometristas, científicos sociales y administradores de la investigación, hemos observado con creciente alarma la omnipresente mala aplicación de los indicadores a la evaluación del rendimiento científico. Los siguientes son solo algunos de los numerosos ejemplos. En todo el mundo, las universidades se han obsesionado con su posición en las clasificaciones mundiales (como el Ranking de Shanghai y la lista de Times Higher Education), incluso cuando dichas listas se basan en lo que, en nuestra opinión, son datos inexactos e indicadores arbitrarios.

Algunos reclutadores solicitan los valores del índice h a los candidatos. Varias universidades basan las decisiones de promoción en los valores del índice h y en el número de artículos en revistas de "alto impacto". Los currículos de los investigadores se

han convertido en oportunidades para presumir de estas puntuaciones, sobre todo en biomedicina. En todas partes, los supervisores piden a los estudiantes de doctorado que publiquen en revistas de alto impacto y adquieran financiación externa antes de estar preparados.

En Escandinavia y China, algunas universidades asignan la financiación de la investigación o las primas en función de una cifra: por ejemplo, calculando las puntuaciones de impacto individuales para asignar "recursos de rendimiento" o dando a los investigadores una prima por una publicación en una revista con un factor de impacto superior a 15 (ref. 2).

En muchos casos, los investigadores y evaluadores siguen ejerciendo un juicio equilibrado. Sin embargo, el abuso de las métricas de investigación se ha extendido demasiado como para ignorarlo.

Por eso presentamos el Manifiesto de Leiden, que lleva el nombre de la conferencia en la que cristalizó (véase http://sti2014.cwts.nl). Sus diez principios no son una novedad para los cienciometristas, aunque ninguno de nosotros sería capaz de recitarlos en su totalidad porque hasta ahora ha faltado la codificación. Algunas personalidades del sector, como Eugene Garfield (fundador del ISI), han declarado algunos de estos principios3,4 . Pero no están en la sala cuando los evaluadores informan a los administradores universitarios que no son expertos en la metodología pertinente. Los científicos que buscan bibliografía con la que rebatir una evaluación encuentran el material disperso en lo que, para ellos, son revistas oscuras a las que no tienen acceso.

Ofrecemos esta destilación de las mejores prácticas en la evaluación de la investigación basada en métricas para que los investigadores puedan pedir cuentas a los evaluadores, y los evaluadores puedan pedir cuentas a sus indicadores.

Crédito: Fuente de datos: Thomson Reuters Web of Science; Análisis: D.H., L.W.

Diez principios

1) La evaluación cuantitativa debe respaldar la evaluación cualitativa de los expertos. Las métricas cuantitativas pueden desafiar las tendencias de sesgo en la revisión por pares y facilitar la deliberación. Esto debería reforzar la revisión por pares, ya que emitir juicios sobre los colegas es difícil sin una serie de información relevante. Sin embargo, los evaluadores no deben caer en la tentación de ceder la toma de decisiones a los números. Los indicadores no deben sustituir al juicio informado. Todos son responsables de sus evaluaciones.
2) 2) Medir el rendimiento en función de las misiones de investigación de la institución, el grupo o el investigador. Los objetivos del programa deben establecerse desde el principio, y los indicadores utilizados para evaluar el rendimiento deben estar claramente relacionados con dichos objetivos. La elección de los indicadores y la forma de utilizarlos deben tener en cuenta los contextos socioeconómicos y culturales más amplios. Los científicos tienen

diversas misiones de investigación. La investigación que hace avanzar las fronteras del conocimiento académico difiere de la que se centra en aportar soluciones a los problemas de la sociedad. La evaluación puede basarse en méritos relevantes para la política, la industria o el público, más que en ideas académicas de excelencia. Ningún modelo de evaluación se aplica a todos los contextos.

3) 3) Proteger la excelencia en la investigación localmente relevante. En muchas partes del mundo, la excelencia en la investigación se equipara a la publicación en inglés. La legislación española, por ejemplo, establece la conveniencia de que los académicos españoles publiquen en revistas de alto impacto. El factor de impacto se calcula para las revistas indexadas en la Web of Science, que tiene su sede en Estados Unidos y sigue siendo mayoritariamente en inglés. Estos sesgos son especialmente problemáticos en las ciencias sociales y las humanidades, en las que la investigación tiene un carácter más regional y nacional. Muchos otros campos tienen una dimensión nacional o regional - por ejemplo, la epidemiología del VIH en el África subsahariana.
4) Este pluralismo y relevancia social tiende a suprimirse para crear artículos de interés para los guardianes del alto impacto: Las revistas en inglés. Los sociólogos españoles que son altamente citados en la Web of Science han trabajado en modelos abstractos o estudian datos estadounidenses. Se pierde la especificidad de los sociólogos en los trabajos de alto impacto en español: temas como la legislación laboral local, la atención sanitaria familiar a los mayores o el empleo de los inmigrantes5. Una métrica basada en la literatura de alta calidad en lengua no inglesa serviría para identificar y premiar la excelencia en la investigación localmente relevante.
5) 4) Mantener los procesos de recogida y análisis de datos abiertos, transparentes y sencillos. La construcción de las bases de datos necesarias para la evaluación debería seguir unas normas claramente establecidas, fijadas antes de que la investigación haya finalizado. Esta fue una práctica común entre los grupos académicos y comerciales que construyeron la metodología de evaluación bibliométrica durante varias décadas. Esos grupos hacían referencia a protocolos publicados en la literatura revisada por pares. Esta transparencia permitió el escrutinio. Por ejemplo, en 2010, el debate público sobre las propiedades técnicas de un importante indicador utilizado por uno de nuestros grupos (el Centro de Estudios de Ciencia y Tecnología de la Universidad de Leiden, en los Países Bajos) condujo a una revisión del cálculo de este indicador6. Los nuevos operadores comerciales deben someterse a las mismas normas; nadie debe aceptar una máquina de evaluación de caja negra.
6) La simplicidad es una virtud en un indicador porque aumenta la transparencia. Pero las métricas simplistas pueden distorsionar el registro (véase el principio 7). Los evaluadores deben esforzarse por lograr un equilibrio: indicadores sencillos y fieles a la complejidad del proceso de investigación.
7) La simplicidad es una virtud en un indicador porque aumenta la transparencia.
8) 5) Permitir que los evaluados verifiquen los datos y el análisis. Para garantizar la calidad de los datos, todos los investigadores incluidos en los estudios

bibliométricos deben poder comprobar que sus resultados se han identificado correctamente. Todos los que dirigen y gestionan los procesos de evaluación deberían garantizar la exactitud de los datos, mediante una autoverificación o una auditoría de terceros. Las universidades podrían implementar esto en sus sistemas de información de investigación y debería ser un principio rector en la selección de los proveedores de estos sistemas. La recopilación y el procesamiento de datos precisos y de alta calidad requieren tiempo y dinero. Hay que presupuestarlo.

9) 6) Tener en cuenta la variación de las prácticas de publicación y citación en cada campo. La mejor práctica es seleccionar un conjunto de posibles indicadores y permitir que los campos elijan entre ellos. Hace unos años, un grupo de historiadores europeos recibió una calificación relativamente baja en una evaluación nacional de revisión por pares porque escribieron libros en lugar de artículos en revistas indexadas por la Web of Science. Los historiadores tuvieron la mala suerte de formar parte de un departamento de psicología. Los historiadores y los científicos sociales exigen que se incluyan los libros y la literatura en lengua nacional en sus recuentos de publicaciones; los informáticos exigen que se cuenten las ponencias en conferencias.

10) Los índices de citación varían según el campo: las revistas de matemáticas mejor clasificadas tienen factores de impacto de alrededor de 3; las revistas de biología celular mejor clasificadas tienen factores de impacto de alrededor de 30. Se necesitan indicadores normalizados, y el método de normalización más sólido se basa en los percentiles: cada artículo se pondera en función del percentil al que pertenece en la distribución de citas de su campo (el 1%, el 10% o el 20% más alto, por ejemplo). Una sola publicación muy citada mejora ligeramente la posición de una universidad en una clasificación basada en indicadores de percentiles, pero puede hacer que la universidad pase de la mitad a la cima de una clasificación basada en promedios de citas7.

7) Basar la evaluación de los investigadores individuales en un juicio cualitativo de su cartera. Cuanto mayor sea la edad, mayor será el índice h, incluso en ausencia de nuevos artículos. El índice h varía según el campo: los científicos de la vida alcanzan un máximo de 200, los físicos de 100 y los científicos sociales de 20 a 30 (ref. 8). Depende de la base de datos: hay investigadores de ciencias de la computación que tienen un índice h de alrededor de 10 en la Web of Science, pero de 20 a 30 en Google Scholar9. Leer y juzgar el trabajo de un investigador es mucho más apropiado que basarse en un número. Incluso cuando se compara un gran número de investigadores, es mejor un enfoque que tenga en cuenta más información sobre los conocimientos, la experiencia, las actividades y la influencia de un individuo.

8) Evitar la concreción errónea y la falsa precisión. Los indicadores de ciencia y tecnología son propensos a la ambigüedad conceptual y a la incertidumbre y requieren fuertes suposiciones que no son universalmente aceptadas. El significado de los recuentos de citas, por ejemplo, ha sido debatido durante mucho tiempo. Por ello, las mejores prácticas utilizan múltiples indicadores para ofrecer una imagen más sólida y plural. Si es posible cuantificar la incertidumbre y el error, por ejemplo

mediante barras de error, esta información debería acompañar a los valores publicados de los indicadores. Si esto no es posible, los productores de indicadores deberían al menos evitar la falsa precisión. Por ejemplo, el factor de impacto de las revistas se publica con tres decimales para evitar empates. Sin embargo, dada la ambigüedad conceptual y la variabilidad aleatoria de los recuentos de citas, no tiene sentido distinguir entre revistas sobre la base de diferencias muy pequeñas en el factor de impacto. Evite la falsa precisión: sólo se justifica un decimal.

9) Reconocer los efectos sistémicos de la evaluación y los indicadores. Los indicadores modifican el sistema a través de los incentivos que establecen. Estos efectos deben anticiparse. Esto significa que siempre es preferible un conjunto de indicadores, ya que uno solo invita al juego y al desplazamiento de objetivos (en el que la medición se convierte en el objetivo). Por ejemplo, en los años 90, Australia financiaba la investigación universitaria mediante una fórmula basada en gran medida en el número de artículos publicados por un instituto. Las universidades podían calcular el "valor" de un artículo en una revista arbitrada; en el año 2000, era de 800 dólares australianos (unos 480 dólares estadounidenses en 2000) en financiación de la investigación. Como era de esperar, el número de artículos publicados por los investigadores australianos aumentó, pero lo hizo en revistas menos citadas, lo que sugiere que la calidad de los artículos disminuyó10.

10) Examinar periódicamente los indicadores y actualizarlos. Las misiones de investigación y los objetivos de evaluación cambian y el propio sistema de investigación evoluciona. Las métricas que antes eran útiles se vuelven inadecuadas y surgen otras nuevas. Los sistemas de indicadores tienen que ser revisados y quizás modificados. Al darse cuenta de los efectos de su fórmula simplista, Australia introdujo en 2010 su iniciativa más compleja Excelencia en la Investigación para Australia, que hace hincapié en la calidad.

Próximos pasos

Respetando estos diez principios, la evaluación de la investigación puede desempeñar un papel importante en el desarrollo de la ciencia y sus interacciones con la sociedad. Las métricas de la investigación pueden proporcionar información crucial que sería difícil de reunir o comprender por medio de la experiencia individual. Pero no se debe permitir que esta información cuantitativa pase de ser un instrumento a ser el objetivo.

Las mejores decisiones se toman combinando estadísticas sólidas con la sensibilidad hacia el objetivo y la naturaleza de la investigación que se evalúa. Se necesitan tanto pruebas cuantitativas como cualitativas; cada una es objetiva a su manera. La toma de decisiones sobre la ciencia debe basarse en procesos de alta calidad informados por los datos de mayor calidad.

Referencias

1. 1

Wouters, P. en Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact (eds Cronin, B. & Sugimoto, C.) 47-66 (MIT Press, 2014).

Google Scholar

2. 2

Shao, J. & Shen, H. Learned Publ. 24, 95-97 (2011).

Artículo Google Scholar

3. 3

Seglen, P. O. Br. Med. J. 314, 498-502 (1997).

Artículo CAS Google Scholar

4. 4

Garfield, E. J. Am. Med. Assoc. 295, 90-93 (2006).

Artículo CAS Google Scholar

5. 5

López Piñeiro, C. & Hicks, D. Res. Eval. 24, 78-89 (2015).

Artículo Google Scholar

6. 6

van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. J. Informetrics 4, 431-435 (2010).

Artículo Google Scholar

7. 7

Waltman, L. et al. J. Am. Soc. Inf. Sci. Technol. 63, 2419-2432 (2012).

Artículo CAS Google Scholar

8. 8

Hirsch, J. E. Proc. Natl Acad. Sci. USA 102, 16569-16572 (2005).

Artículo ADS CAS Google Scholar

9. 9

Bar-Ilan, J. Scientometrics 74, 257-271 (2008).

Artículo CAS Google Scholar

10. 10

Butler, L. Res. Policy 32, 143-155 (2003).

Artículo Google Scholar

Descargar referencias